Gender and Age Estimation from Synthetic Face Images with Hierarchical Slow Feature Analysis

Alberto N. Escalante B. and Laurenz Wiskott

Institut für Neuroinformatik, Ruhr-University of Bochum, Germany,
{alberto.escalante|laurenz.wiskott}@ini.rub.de

Abstract. Our ability to recognize the gender and estimate the age of people around us is crucial for our social development and interactions. In this paper, we investigate how to use Slow Feature Analysis (SFA) to estimate gender and age from synthetic face images. SFA is a versatile unsupervised learning algorithm that extracts slowly varying features from a multidimensional signal. To process very high-dimensional data, such as images, SFA can be applied hierarchically. The key idea here is to construct the training signal such that the parameters of interest, namely gender and age, vary slowly. This makes the labelling of the data implicit in the training signal and permits the use of the unsupervised algorithm in a hierarchical fashion. A simple supervised step at the very end is then sufficient to extract gender and age with high reliability. Gender was estimated with a very high accuracy, and age had an RMSE of 3.8 years for test images.

Keywords: Slow feature analysis, human face images, age, gender, hierarchical network, feature extraction, pattern recognition.

1 Introduction

The estimation of gender and age is crucial for many social interactions, and is done everyday consciously or unconsciously. This process happens very quickly and requires relatively little visual information which is usually of dynamic nature, but we are also capable of performing this process with still images.

In this work we investigate how an unsupervised algorithm for signal extraction can be used to automatically extract gender and age information from single frontal images of simulated subjects (random 3D face models). This has applications to man-machine interaction, face recognition, and as an aid in the supervision of age and gender related policies.

In order to learn the gender and age of the subjects, we decided to use a versatile unsupervised algorithm called Slow Feature Analysis (SFA). SFA extracts slowly varying features from a high-dimensional signal. Contrary to other unsupervised learning algorithms, for SFA time plays a key role. In this paper, the high-dimensional signal is a sequence of images (e.g. each image is a $135 \times 135 = 18225$-dimensional vector), and it is enforced that one or more (hidden) parameters involved in image generation change on a relatively slow timescale. Although individual signal components (e.g. pixels of an image) might change on a very fast timescale, the algorithm should find a way to combine several signal components at any time step, such that the resulting computed signals vary each as slowly as possible over time, while still containing information about the input.

The trick in using this unsupervised algorithm to learn some particular feature is to create an appropriate training sequence in which the slowest varying parameter is the feature we want to learn. Thus, for instance, for the age estimation problem, the training signal is a sequence of face images in which the age of the subjects increases very slowly. We show that in this case, the slowest learned feature is strongly correlated with the original age of the subject.

1.1 Related Work

Berkes et al. [3, 4] used (a single unit of) quadratic SFA to analyze sequences of image patches from natural images. They studied optimal stimuli yielding the largest and smallest responses from the unit, and showed that SFA is capable of learning receptive fields with properties similar to those found in complex cells in the primary visual cortex.

Later Franzius et al. [5] implemented a hierarchical model of SFA and used it to extract position and view direction in a simulated box environment. They showed that the type of features learned, which resemble certain cells in a rodent's brain, depend solely on the statistics of the sequences of images generated by the movement inside the box.

More recently, Franzius et al. [6] also used the temporal slowness principle and followed an invariant object recognition approach. They estimate the identity and pose of artificial fish and textured spheres from still images. They studied the simultaneous change in one or more slow parameters at different timescales. Contrary to this work, the supervised post-processing used for feature estimation is based on linear regression and they used a much larger number of signals for this step, while we used only three signals.

Some existing methods for gender classification, which can be roughly divided into appearance-based and geometric-based approaches, are briefly described in [9, 7] and for age classification in [7].

2 Slow Feature Analysis (SFA)

SFA is a biologically inspired unsupervised learning algorithm [8], that in its linear version is somewhat related to PCA and ICA, but has the essential property that the temporal component of the variables is also considered (i.e., the temporal ordering of the samples matters).

The input is a multidimensional signal $\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))^T$. SFA then computes a set of weights $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,N})^T$, such that each output signal

$y_i(t) = \boldsymbol{x}(t)^T \boldsymbol{w}_i$ has the slowest possible temporal variation and is uncorrelated to signals y_j for $j < i$.

More formally, the output signals $y_i(t)$, for $0 \leq i < N$ must be optimally slow in the sense that the objective function $\Delta(y_i) \overset{\text{def}}{=} \langle \dot{y}_i(t)^2 \rangle$ (i.e., the variance of the time derivative of y_i) is minimal while the following constraints must hold:

- Zero mean: $\langle y_i(t) \rangle = 0$
- Unit variance: $\langle y_i(t)^2 \rangle = 1$
- Decorrelation: $\langle y_i(t) y_j(t) \rangle = 0$ for $j < i$

The SFA problem is to find an optimal set of weights $\{\boldsymbol{w}_i\}$ such that the conditions above are met. Fortunately, it is well known that the optimal solutions to this problem depend only on the covariance matrix $B = \langle \boldsymbol{x}\boldsymbol{x}^T \rangle$ of the training sequence $\boldsymbol{x}(t)$, and the covariance matrix $A = \langle \dot{\boldsymbol{x}}\dot{\boldsymbol{x}}^T \rangle$ of the time derivative of the training sequence $\dot{\boldsymbol{x}}(t)$. In practice, time is discrete and the time derivative is approximated by the difference of consecutive samples in the training sequence.

Moreover, it is possible to state the SFA problem as a generalized eigenvalue problem, and traditional algorithms for solving the latter problem can be used. As a consequence, the algorithm has a similar complexity as PCA and is guaranteed to find an optimal solution.

3 Hierarchical Slow Feature Analysis

To apply even linear SFA on the whole training data would be too expensive, since it would have a computational complexity of $\mathcal{O}(LN^2 + N^3)$ where L is the number of samples and N is the dimensionality. This complexity problem becomes more severe if a non-linear preprocessing step is applied to the images to obtain non-linear SFA. Hierarchical SFA allows us to cope with this problem by dividing the image sequence in smaller dimensionality sequences that are separately processed by SFA units, cf. [5]. Afterwards, the slow signals separately computed by these units can be grouped together and further processed by the SFA units in the next layer. This process can be repeated and organized in a multi-layer hierarchy until global slow features are extracted, where each layer is trained separately from the first to the last.

Although hierarchical networks based on SFA have been successfully tested on different stimuli before, e.g. images of fish, textured spheres [6] and the interior of boxes [5], it is unclear whether this type of network would also succeed at learning from frontal face images, because changes in the slow parameters in the training data only produce subtle changes at the pixel level (compared for example to fish identity or pose, which offer larger variability at the pixel level). We prove that hierarchical SFA is powerful enough to learn these slow parameters.

The hierarchical SFA networks we have developed can be employed unchanged to extract different relevant parameters from image sequences, where the learned parameters are implicit in the training data. Thus, we only need to modify the training set according to the particular parameter to be learned.

A special effort was made to keep the computational cost of the training procedure low because, as in many learning algorithms, training is a relatively expensive procedure. However, once trained the computational and memory cost for SFA are very low, and thus feature extraction from a single image is a fast procedure.

We built several networks and tested several values of the parameters that define its structure and the composition of the layers. In this article, we focus only on one particular linear and a non-linear network. We remark that its structure is not problem-specific, except for the input dimensionality of the networks which should agree with the image size. This is in accordance with the desire of building a flexible architecture capable of tackling different problems without modification.

Linear SFA Network This is the simplest network we developed. As any linear system, it has well known limitations that reduce the type of relevant features that can be correctly extracted. This limitation is slightly reduced by the use of a Gaussian classifier on top of the linear network (see Section 4.3 on post-processing).

The network has 4 processing layers, which operate one after the other and reduce the dimensionality from 135x135 pixel values in the input images to just 40 signals at the network output. Each layer can be further subdivided into a few elementary sub-layers, which in turn are composed of elementary data processing units arranged in a square array. These units can be, for example, SFA nodes or PCA/whitening nodes. For reasons of space we omit here the details of the network structure. The first layer contains an SFA sub-layer with 27x27 SFA nodes, each one having a non-overlapping fan-in of 5x5x(1 pixel) and a fan-out of 16 signals, thus reducing the data dimensionality by 36%. Similarly, the second layer has a 9x9 grid structure, each unit has a fan-in of 3x3x(16 signals) and a fan-out of 30 signals, which reduces the data dimensions from 27x27x16 to 9x9x30 signals, a further reduction of 79%. In the same way, the third layer has a 3x3 grid structure, each unit has a fan-in of 3x3x(30 signals) and a fan-out of 40 signals. The forth layer has a single SFA node that takes the whole output of the previous layer and outputs only 40 signals. The complete network reduces the amount of signals from 135x135 to just 40 signals, where only 3 of them are given to the classifier.

Non-Linear SFA Network Our non-linear network has the same architecture as the linear network, with the only difference that non-linear expansion nodes are added in each sub-layer before the SFA nodes. These nodes introduce some amount of non-linearity that depends on the expansion function that was chosen. The more powerful this expansion is, the more capable the network becomes in extracting complex features. Therefore, it is tempting to use a complex expansion, say a 5th degree product expansion, where all products up to degree five on the input signal components appear. However, a large expansion increases the computational cost and the amount of training data needed to avoid overfitting.

Therefore, more conservative non-linearities are typically preferred, such as a quadratic expansion (including all terms of the form x_i and $x_i x_j$ for $0 \leq i, j < N$, where $(x_0, \ldots, x_{N-1})$ is the original signal). In this work, we use modest non-linearities. The expansion function computes all terms $x_i x_{i+1}$ for $0 \leq i < N-1$ in addition to the linear terms x_i. Each non-linear expansion node roughly doubles the number of signals. However, the number of slow signals extracted by the SFA nodes is kept the same as in the linear case to avoid an explosion in the number of signals.

Other expansions that we have tested include the product of pairs of variables with similar slowness values, and sums or differences instead of products combined with other non-linearities such as absolute values and square roots. We did not find any advantage in using these expansions.

4 Training and Test Sequences for Age and Gender Estimation

After having built suitable SFA networks, the next step was to generate an appropriate data set for training and testing. The network learns to estimate gender or age from artificial frontal images based solely on the particular sequence of images used for training. After training we separately test its performance with respect to these images and new images not seen before by the network. All the training and test images were generated in software only once before training took place.

The software used for face model generation is called FaceGen [2], image rendering was done with POV-Ray [1], other tools were used for format conversions, and the process was partially automated with many Perl scripts, and a few Python scripts. The arguably large amount of images is required to reduce overfitting.

4.1 Sequences for Gender Estimation

The first data set was created as follows. A large number of random subjects was needed. In this case, we created 12000 random subjects, each one defined by a unique 3D face model without hair, glasses, earrings or other accessories. These models are generated with several randomized low- and high-level facial parameters that include (at a high-level) age, symmetry, gender and racial composition, and it is possible to change any of these parameters. For example, the gender parameter is a real value, defined by the software for face generation as: -3 = very masculine, -1=masculine, 1=feminine to 3 = very feminine. This allowed us to arbitrarily select the level of masculinity or femininity of the models, and thus create sequences of images of random subjects where the gender value slowly increases from very masculine to very feminine. We selected 60 fixed gender values: $(-3, -2.9, \ldots, 2.9)$ and 200 subjects per gender value, thus requiring 12000 face images. A neutral expression was chosen, random vertical and horizontal translations of +/- 2 pixels were added to each image, and pink-noise like random backgrounds were used. Notice that the addition of a translation and randomized backgrounds makes the problem more difficult and is inspired by more realistic conditions of real photographs. The network should now learn to remain invariant to small translations. It should actually also become invariant to the quickly changing randomized background since it is not a good source of slow signals.

The training sequence (Figure 1) is composed of 180 of the subjects for each gender value accounting for 10800 images, while the test sequence is composed of the remaining 20 subjects per gender value accounting for 1200 images.

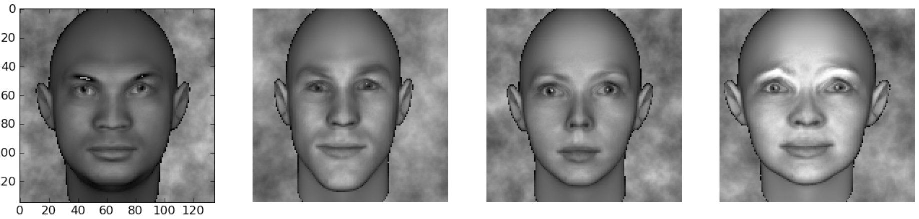

Fig. 1. A few examples of the images of the training sequence used for gender estimation. The gender parameter varies here from -3.0 (left), -1.1, 0.9 to 2.9 (right).

4.2 Sequences for Age Extraction

The face generation software only allows for generating subjects from 15 to 65 years. For efficiency purposes, we selected 23 specific ages non-uniformly, increasing from 15 to 65 years: (15, 16, 18, 20, …, 55, 60, 65 years). The separation between samples was shorter for smaller ages because we expected a larger variability at the pixel level in young subjects than in older subjects.

We created 200 random subjects for each age value, accounting for 4600 random subjects of different ages. Again, no hair, glasses, earrings or other accessories were present. Also a neutral expression was chosen, pink-noise like random backgrounds were used, and smaller random vertical and horizontal translations of +/- 1 pixel were added to each centered image.

For the training set we took 180 of the generated subjects for each age value accounting for 4140 images, while the test sequence is composed of the remaining 460 images.

4.3 Supervised Post-Processing of the Slow Signals

A classifier is taught to relate the output of the network to the known values of the relevant parameters, such as the true age or gender of the input samples (while the network itself is unsupervised, the labels with the known gender or

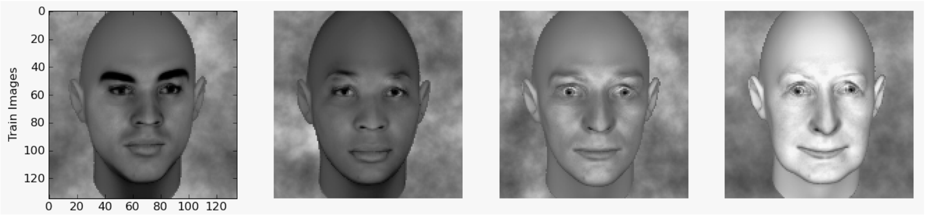

Fig. 2. Examples of the images of the training sequence used for age estimation. The age parameter varies here from left to right from 15, 26, 44 to 65 years.

age are used to train the classifier). For the linear network, this constitutes the single non-linear step in the architecture.

Theoretically, we expect that the slow signals extracted by the network should depend on the slow parameter that we want the network to learn. Notice however, that the slowest signal does not have to be linearly related to the slow parameter, so it might not be possible to use it directly to recover the parameter. What we need is a way to establish a connection between the domain of the slow signals, and the parameter domain. The classifier takes advantage of the fact that the slow parameter is redundantly coded in the slow signals, as the theory indicates, as the slowest signal and as its harmonics. Additionally, we exploit the fact that the training set is labelled (since we know gender and age during image generation) to estimate the parameter. In theory, images with the same slow parameter cluster in a single point in the output domain. We use a small set of slow signals, here the 3 slowest output signals, to train a classifier. As labels for the classifier we use the real gender or age parameter. If the network generalizes well, then the classifier should be able to output the correct value of the parameter for new images. Moreover, if class probabilities are present, we can improve the estimation of the parameter aiming at minimizing the MSE. Two classifiers were used: a closest center classifier and a Gaussian classifier. A class was defined for each possible value of the labels.

We assumed that the Gaussian Classifier perfectly learned the distribution of the data, and is able to perfectly estimate the class probabilities. Then, we used the class probabilities and the labels to find the value that minimizes the MSE. Let $P(l_i)$ be the probability that a given image actually has label l_i, for $1 \leq i \leq C$, then our estimate of the parameter is $\sum_i l_i P(l_i)$, where i ranges over the C classes.

5 Results

For the gender extraction experiment, the linear SFA network followed by a simple Gaussian classifier on 3-dimensional signals was capable of estimating the gender of new random subjects with a root mean squared error (RMSE) of 0.33 (Table 1). Recall that the gender parameter varies in the interval $(-3, 2.9)$.

Thus the standard error from the true parameter is about 5% of the parameter's range.

	Linear Network		Non-Linear Network	
	RMSE Gaussian	RMSE CCC	RMSE Gaussian	RMSE CCC
Gender Estimation				
Training Images	0.3084	0.6098	0.2229	0.3119
Test Images	0.3318	0.6857	0.4180	0.5858
Age Estimation (years)				
Training Images	3.2963	5.6461	2.2313	3.3914
Test Images	3.8365	7.0697	5.3618	7.9660

Table 1. Performance of the networks in terms of the root mean squared error (RMSE) using a Gaussian classifier (GC) and a closest center classifier (CCC).

In Figure 5 we can see the three slowest signals extracted by the linear network from the training sequence of the gender estimation experiment. Notice that the slowest signal (in black) is less noisy than the other signals. The same figure for the test sequence (not shown) is very similar, except that it has fewer data points.

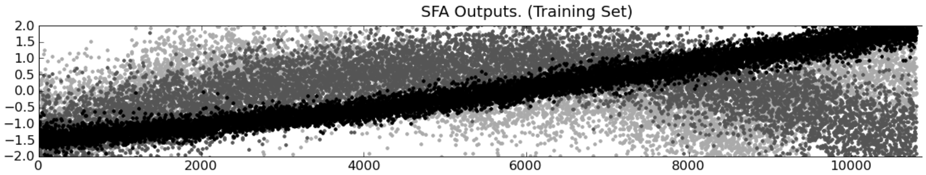

Fig. 3. Linear network: slowest signals for the gender experiment. The black, dark grey and light grey points are the slowest, second slowest and third slowest signals, resp. The horizontal axis indicates the image number, and the vertical axis is the amplitude of the slow signals.

The reported performance was achieved when the classifier was trained with only the 3 slowest signals computed by the network. The precise number of signals given to the classifier has a direct influence on the performance of the system. Its optimal value depends on the combination of the network employed and the training sequence.

Using only one signal degrades the quality of the estimation, because it reduces the available redundancy. Using many signals, however, is not useful because faster varying extracted signals are increasingly noisier than the slowest ones, thus the classifier cannot take much advantage of them. Moreover, if the number of signals is increased, the Gaussian classifier also needs more samples to reliably learn the input distribution.

The non-linear network performs better on the training data than the linear one, as expected, but suffers from more overfitting, which explains why it does not outperform the linear network on new data. The non-linear network will become superior for newer data once enough training samples are used.

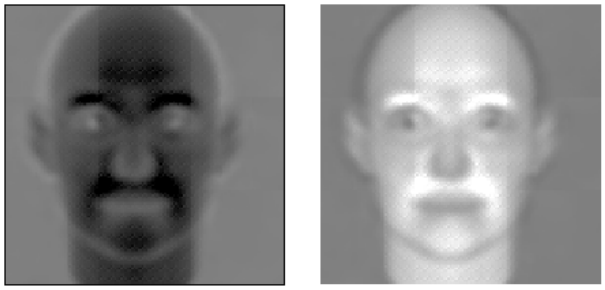

Fig. 4. The linear mask (weights) that encodes the slowest output signal and its negative (normalized for display purposes). Notice how the first image resembles more a masculine face, while the second a feminine one.

The problem of age estimation is more difficult than gender estimation. In informal tests, it was clear that the ability of a human operator at estimating age from the images was limited. Thus we were not expecting a good performance from the system. The linear network had an RMSE of 3.8 years from the true age of the subjects, and 3.3 years for the training samples. The performance of the non-linear network for the training samples was clearly superior with an RMSE of only 2.2 years. Unfortunately, again it did not generalize as well because we did not use enough samples.

6 Conclusion and Future Work

We developed two very flexible hierarchical networks for slow feature analysis. These networks are application independent, and can be used to learn slow parameters from very different two-dimensional signals. Training was accomplished in less than 30 minutes. Importantly, the output of the network agreed to a large extent with the theoretically predicted properties of SFA on the whole images.

The expansion of the data in a non-linear way, even a small expansion, increases the performance of the network, but has the disadvantage that larger training sequences are required, otherwise the generalization property is diminished. The amount of training data was earlier shown to be related to the number of features that the system must become invariant to. Hence the addition of rotation, translation, scaling, glasses, clothes, etc. require more training data for the network to be able to ignore such features.

It must be underlined that the networks learn slowly varying parameters according to the underlying model used by the face generation software. Learning from real face images is an interesting topic that we are currently studying. For age and gender estimation using normalized real images we expect a small decrease in the performance. The development of a full SFA-based pipeline for face detection, pose estimation and face recognition is also a challenging topic that we would like to address.

As future work, we will develop more complex SFA hierarchies and design methods to reduce the amount of training data and specially labelled data required, which is now the main factor required to handle real images with this architecture.

Acknowledgments

We would like to thank Marco Müller for his support in using the software for face generation, and Michael Brauckmann for inspiring discussions and for his technical support.

References

1. POV-Team, POV-Ray. `http://www.povray.org`.
2. Singular Inversions Inc., FaceGen SDK. `http://www.facegen.com`.
3. P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.*, 5(6):579–602, 7 2005.
4. A. Escalante and L. Wiskott. Gender and age estimation from synthetic face images with hierarchical slow feature analysis. In E. Hüllermeier and R. Kruse, editors, *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2010*, 2010.
5. M. Franzius, H. Sprekeler, and L. Wiskott. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, 3(8):e166, August 2007.
6. M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition with slow feature analysis. In V. Kurkov, R. Neruda, and J. Koutnk, editors, *Proc. 18th Intl. Conf. on Artificial Neural Networks, ICANN'08, Prague*, volume 5163 of *Lecture Notes in Computer Science*, pages 961–970. Springer, Sept. 2008.
7. V. K. R. Ramesha K, K B Raja and L. M. Patnaik. Feature extraction based face recognition, gender and age classification. *International Journal on Computer Science and Engineering (IJCSE)*, 02(01S):14–23, 2010.
8. L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.
9. X.-C. L. Zheng Ji and B.-L. Lu. *State of the Art in Face Recognition*, volume 5507/2009 of *Lecture Notes in Computer Science*, chapter Gender Classification by Combining Facial and Hair Information, pages 647–654. Springer Berlin / Heidelberg, July 2009.